# キー概念にもとづく情報検索と検索結果の順位付けの検討

3V－10

藤崎博也　大野澄雄　阿部賢司　飯島岐勇　片見憲次　鈴木匡芳

東京理科大学

## 1. はじめに

情報検索においては、多くの場合、ユーザは検索すべき対象を当初から明確に意識しているわけではなく、また、十分な知識を持ち合わせているとは限らない。人間が介入する検索においては、検索の専門家(サーチャー)がユーザとの音声による対話を通じてユーザの検索要求を明確化させ、迅速かつ的確な検索を行なっている。したがって、機械による情報検索システムにおいても、ユーザとの音声対話によりその検索要求を明確にすることが望ましいと考えられる。また、キーワードによる従来の情報検索では、語の表記のみに着目して処理するため、異表記同義・同表記異義の存在が検索性能の低下をもたらす。これを避けるには、キー概念を用いることが有効である[1]。このような見地から、筆者らは既に音声対話とキー概念検索、さらに未知語処理・知識獲得・エージェント技術を組み合わせた、新しい情報検索システムの構想を示している[2]。

本稿では、このシステムの具体化を目的として、学術情報検索における対話の実例を収集、分析し、また、キー概念にもとづくキーワードの集合の拡張を行い、さらに適切なものが上位に集中するような検索結果の順位付けの方法について検討する。

## 2. 対話例の収集と分析

前節の目的のための基礎データとして、学術情報検索を目的とした人間対人間の対話49例を収集した。この場合の対話例を分析すると、その流れは図1のように整理できる。

一方、以下で述べるキーワードの自動抽出の手法を検討するために、検索式を生成することに熟練した者が個々の対話例を見て、キーワード(以下、"キーワードA"と呼ぶ)になると考えた語を選定し、それ用いて最適と考える検索式を生成した。以下、これを"検索式A"と呼び、その検索結果を"結果A"と呼ぶ。

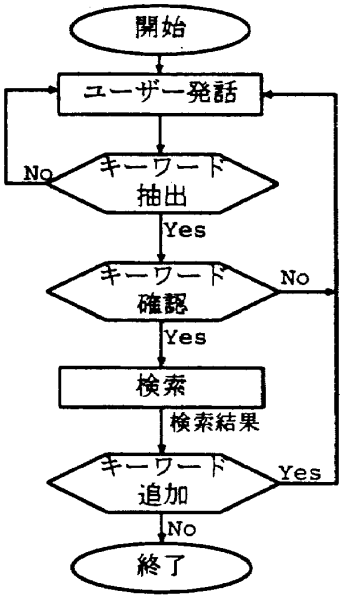


図1. 対話の流れ

さらに、このようにして選定されたキーワードAの、対話の中で占める文法的役割を調べるため、形態素解析ツール「茶筌」[3]を用いて分析し、その結果を分類した(表1)。

表1 キーワードの分類

| 分類 | 割合 |
|---|---|
| 複数の名詞からなる語 | 47.6％ |
| 単独の固有名詞からなる語 | 20.6％ |
| 単独の普通名詞からなる語 | 19.9％ |
| 形容詞と名詞からなる語 | 2.4％ |
| 記号、数詞、名詞からなる語 | 2.4％ |
| 名詞と接尾語からなる語 | 0.8％ |
| 人名がうまく分けられなかった語 | 4.8％ |
| その他 | 0.8％ |

## 3. 検索の手順

### 3.1 キーワード候補の抽出

前節の分析を踏まえ、以下の条件を満たす語をキーワード候補として抽出する。

(1) 1個、または2個以上連続する名詞からなる語

(2) 形容詞とそれに続く1個、または2個以上の名詞からなる語

Information retrieval and ranking of retrieved results based on key concepts
Hiroya Fujisaki, Sumio Ohno, Kenji Abe, Michio Iijima, Kenji Katami, and Masayoshi Suzuki
Science University of Tokyo, 2641 Yamazaki, Noda, 278-8510

(3) 連続する記号、数詞、名詞からなる語
(4) 名詞と接尾語からなる語

### 3.2 検索不要語の排除

抽出されたキーワード候補の中には、キーワードとして不適切な語 (以下、検索不要語と呼ぶ) も含まれている。これらは a) 検索対象の特徴付けを行うには不適切な語、b) キーワードになる場合もあるが当該の対話では不適切な語、c) 形態素解析の誤りによって抽出された文字列、の３種類に分類することができる。これらは以下の方法にしたがって排除する。

(1) a), c) の場合

予め作成した不要名詞リストによって排除する。

(2) b) の場合

(2-1) サ変複合動詞の一部の場合

直後にサ変動詞「する」の活用形があり、かつ不要動詞リストにある語は排除する。

(2-2) 特別な語の場合

例えば「自然言語処理関係」が抽出された場合、キーワードとするのは「自然言語処理」という部分だけである。このことから連続する名詞の最後尾に現れた「関係」は検索不要語とする。このように、特別な語ごとに規則を作り、対処する。

### 3.3 キーワード集合の拡張と検索式の自動生成

抽出したキーワードをＥＤＲ電子化辞書を用い、キー概念のレベルまで遡り、同じ概念を持つ他の語もキーワードであるとみなしてキーワードの集合を拡張する。概念が複数存在する (同表記異義) 場合には、ユーザに質問し、概念を特定する。このようにして得られたキーワードを全て OR 演算子で結んだ検索式 ( 以下、"検索式 B" と呼ぶ) を生成する。

### 3.4 検索式 B による検索結果の順位付け

検索式 B では検索洩れを防止することはできるが、不要な検索は避けられないため、検索式 B の検索結果 (以下、"結果 B" と呼ぶ) は膨大な数になり、検索効率の観点からはよい検索式とは言えない。一方、検索式 A は理想に近いと考えられるため、ここでは結果 A を便宜上の目標とする。したがって、結果 B の上位に目標が集中するように、適切な順位付けを行うことが望ましい。そのような方法の一つとして以下の規則による定量的基準を用いた順位付けを試みる。

(1) キーワード出現位置による得点

キーワードが文献のタイトルやキーワード、著者名に出現している場合、大きな得点を加える。

(2) 同一キーワードの出現回数による得点

同一のキーワードが一つの文献に複数回現れる場合、２回目以降は加える得点を小さくする。

## 4. 順位付けの評価

収集した対話による検索 85 件それぞれについて、結果 B の一定の割合以内に結果 A の全ての文献が現れている検索の件数を調べた。また、結果 A の文献のうちの 50%についても同様に調べ、その件数を示した (表 2(a))。一方、結果 B の上位 10 位までの中に含まれている、結果A以外の文献の個数 (目標外数) 別に検索の件数を示した (表 2(b))。

表 2 順位付けによる効果

(a) 結果 A が結果 B の各割合以内にある検索の件数

| 対象 | 10% | 20% | 30% | 40% | 50% | それ以上 |
|---|---|---|---|---|---|---|
| 結果 A の全て | 77 | 2 | 0 | 2 | 2 | 2 |
| 結果 A の 50% | 83 | 0 | 0 | 1 | 1 | 0 |

(b) 結果 B の上位 10 位までに含まれる目標外数別の検索の件数

| 目標外数 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 件数 | 64 | 2 | 0 | 0 | 1 | 0 | 4 | 1 | 0 | 8 | 5 |

表 2(a), (b) は、結果 B を順位付けすることにより、結果 A の大部分が結果 B の上位に含まれることを示している。なお、表 2(b) において目標外に分類されたものの中に、検索意図に合致する文献が現れた例も、極めて少数 (85 件中 3 件) ではあるが存在した。

## 5. おわりに

本稿では、キー概念にもとづきキーワードの集合を拡張することにより、検索式を生成し、得られた検索結果に適切な順位付けを行う方法を検討した。今後はさらに、不要な検索を減らす方法を検討する。

### 参考文献

[1] 藤崎博也, 亀田弘之, 河井恒 : "新聞記事情報の階層構造に基づく記事分類・検索システム," 情報処理学会「自然言語処理」研究会資料 44 - 4 (1984).

[2] 藤崎博也, 亀田弘之, 大野澄雄, 阿部賢司, 伊東卓哉, 佐久間聖仁 : "キー概念の抽出と未知語の処理に基づく情報検索方式の高度化," 情報処理学会第 54 回全国大会講演論文集, vol. 3, pp. 23-24 (1997).

[3] 松本裕治, 北内啓, 山下達雄, 平野善隆, 今一修, 今村友明 : "日本語形態素解析システム「茶筌」 version 1.5 使用説明書," Technical Report NAIST-IS-TR97007(1997).